

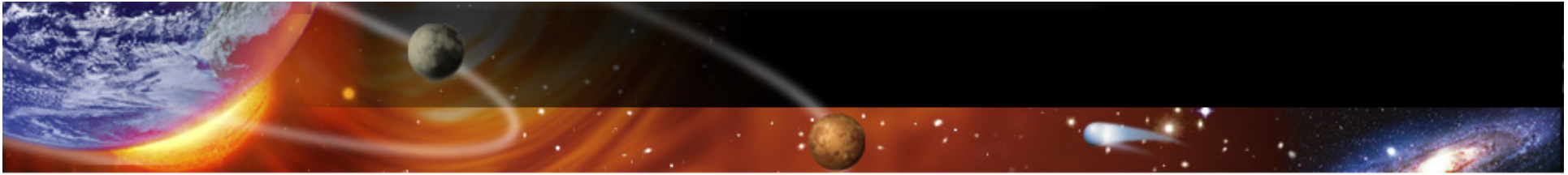


Multi-Scale Structure in 3-D Surveys and Simulations of the Universe

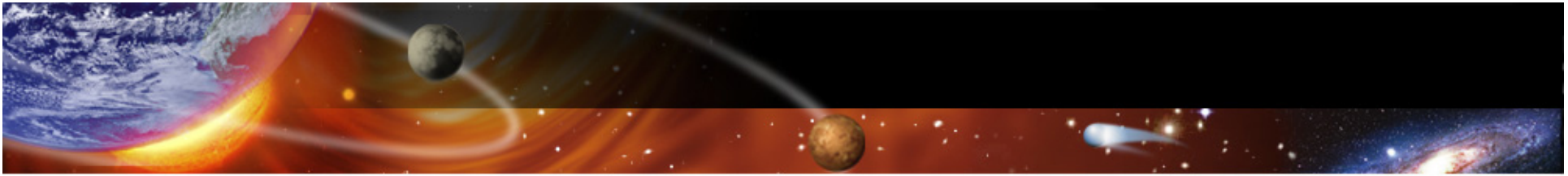
Michael Way (NASA/Goddard Institute for Space Studies)

Paul Gazis, Jeffrey Scargle (NASA/Ames, Space Sciences Division)

<http://astrophysics.arc.nasa.gov/~mway/lss201010.pdf>



- Three structure analysis methods:
 - Kernel Density Estimation
 - Bayesian Blocks
 - Self-Organizing Maps
- Three data sets:
 - Sloan Digital Sky Survey DR7
 - Millennium Simulation
 - Random/Uniform/Independent “Poisson”



Three cornerstones of Data Mining and Machine Learning

Three Steps

- Points → → → Density Estimate
- Density Field → → → Cluster Identification
- Clusters → → → Classification

- Why do we care about the 3-D structure of our universe?
- Building a 3-D catalog: SDSS, MS, Uniform
- Three density estimation methods
- Results



Who cares?

- 1) Astronomers use n-body models to describe the evolution of structure of the universe
- 2) These models must be compared with the real world
- 3) We must find ways to characterize and then compare the structures in each
- 4) This should lead to better models, or at least embarrassment!



3-D Catalogs

The SDSS Volume Limited Catalog

- 1.) “MAIN” sample galaxies from the SDSS DR7: 561,421
- 2.) Make a geographically contiguous region in RA/DEC around the NGP
- 3.) Make a conservative Volume Limited cut: $z < 0.12$ $R < -20.07$
- 4.) End up with 114,112 objects

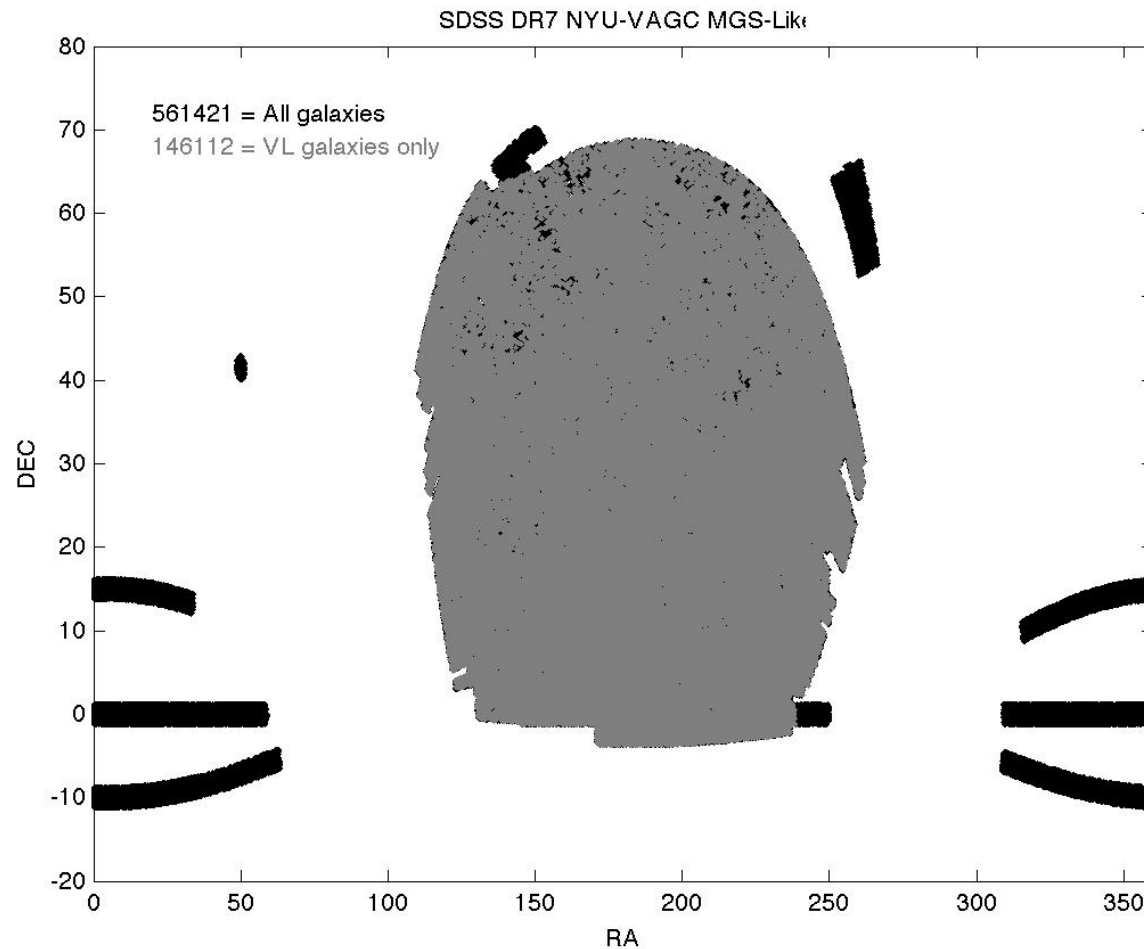


3-D Catalogs

- 5.) The fiber positioners of the SDSS are limited to $> 55''$
- 6.) There is plate overlap, so to make sure our sample is chosen homogeneously eliminate any galaxy within $55''$ of another
- 7.) Boundary point problem: Voronoi Tessellation will give extremely large-volume/low-density boundary cells. These must also be eliminated: 133,991 left over.
- 8.) Bright Stars? Not believed to be a problem ($0 < m_v < 6.5$)

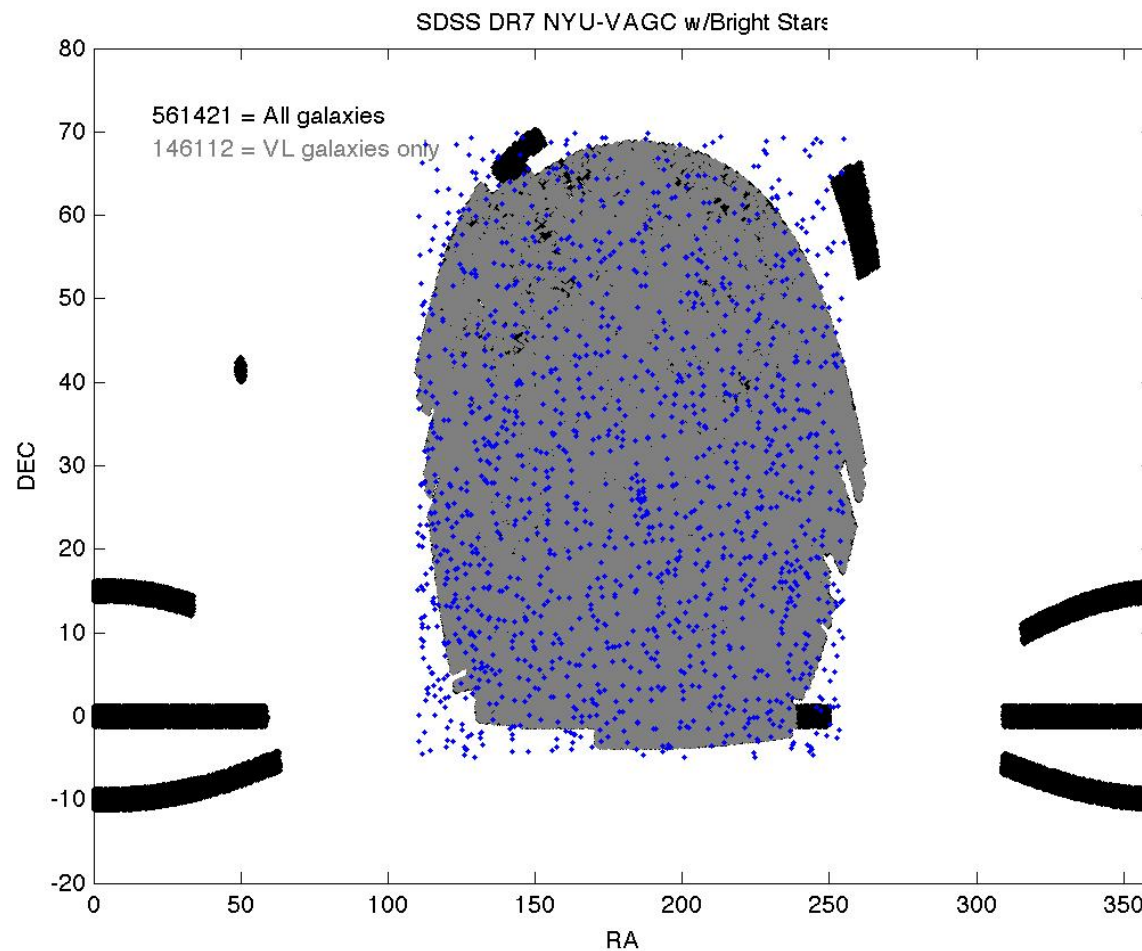
The SDSS DR7 NYU-VAGC

The SDSS Data Release 7 “MAIN” Galaxy Sample



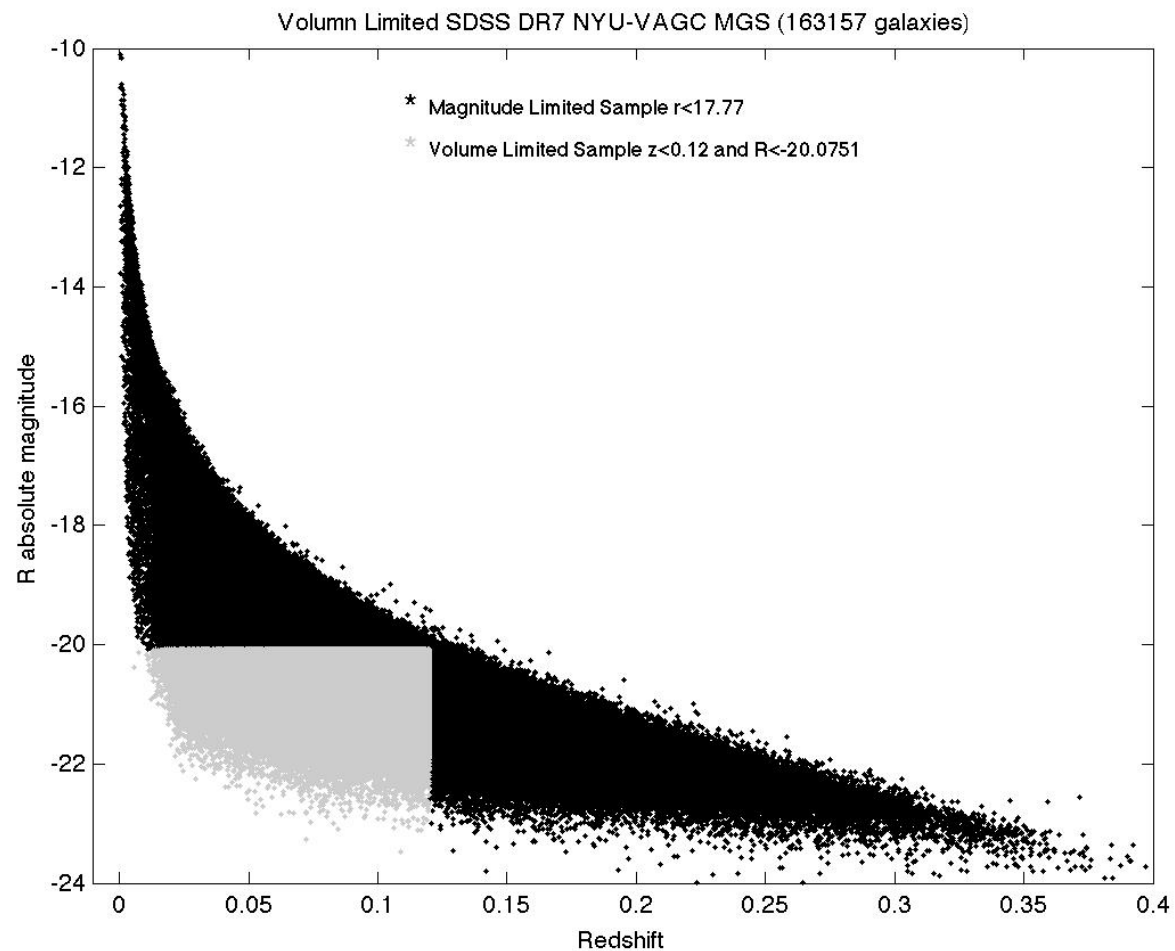
The SDSS DR7 NYU-VAGC

with 5th revised ed. Bright Star Catalog ($0 < m_v < 6.5$)



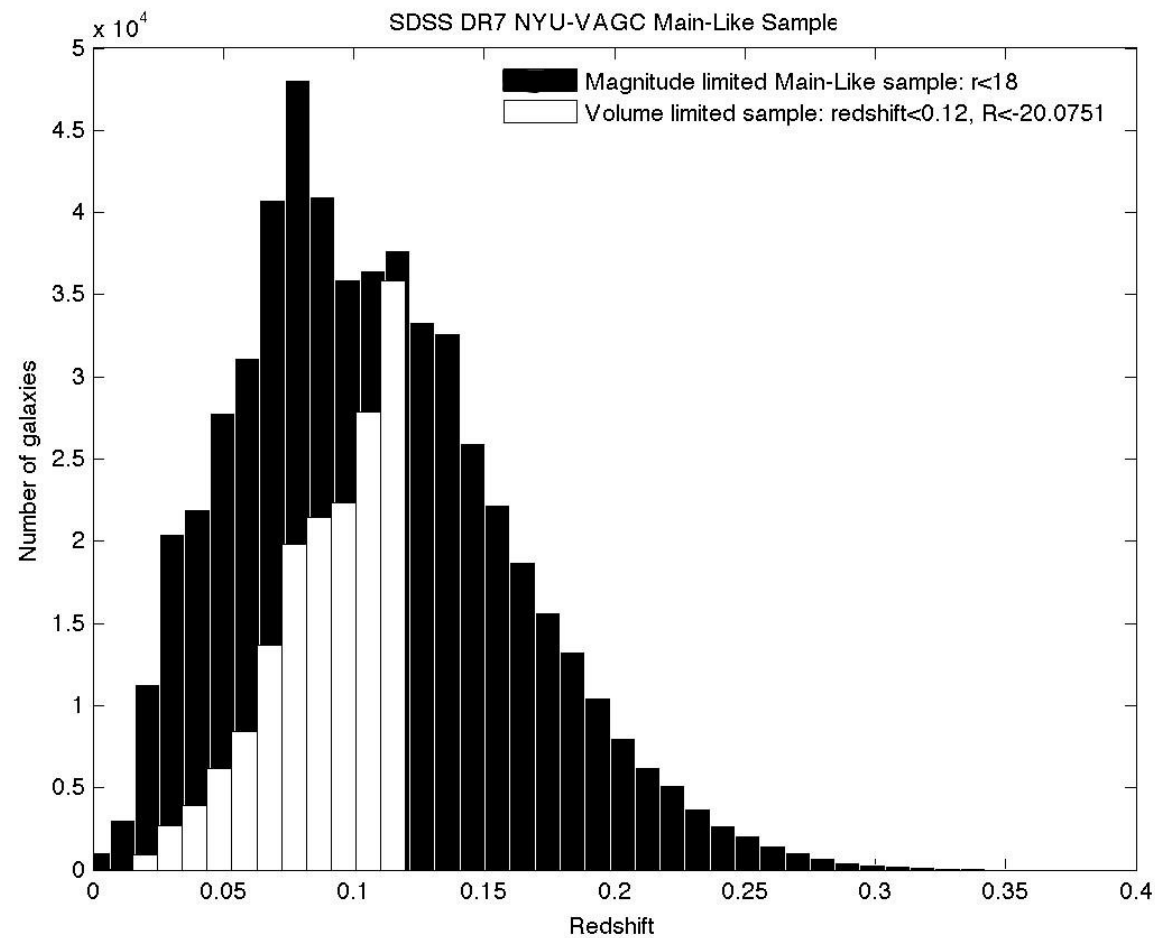
SDSS DR7 Vol. Limited

Picking the volume limited sample



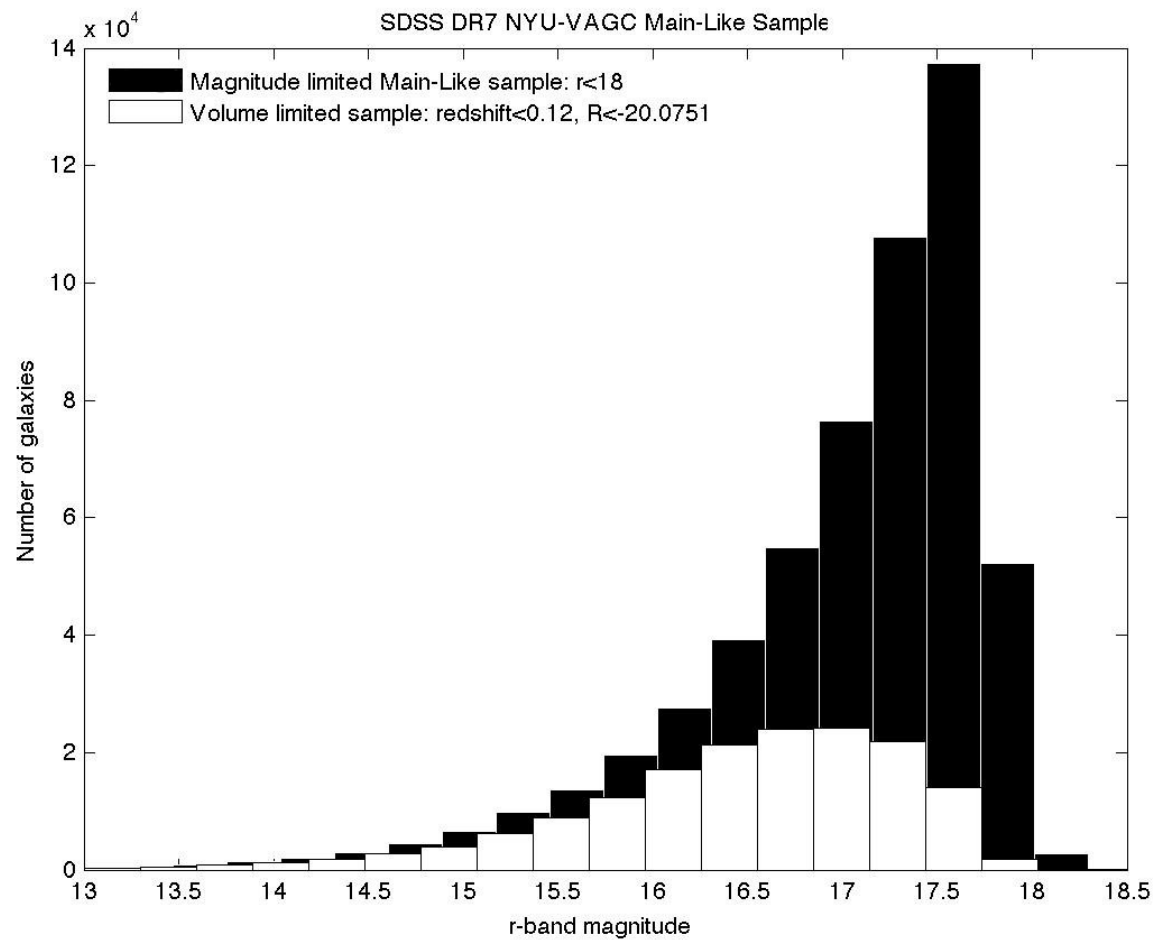
SDSS DR7 Vol. Limited

Picking the volume limited sample



SDSS DR7 Vol. Limited

Picking the volume limited sample





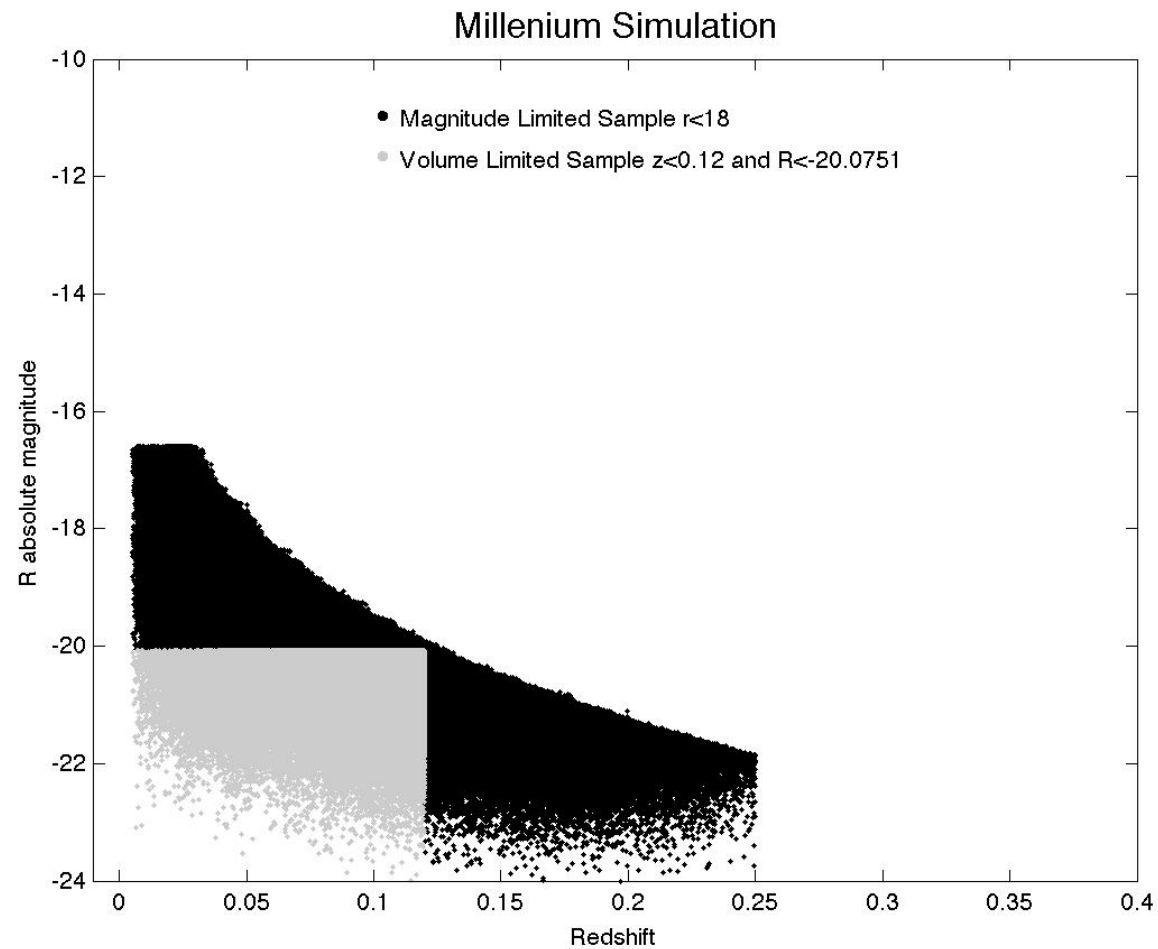
3-D Catalogs

Millennium Simulation & Poisson Catalogs

- The same methods used to create the SDSS catalog are used to derive a similar catalog from the Millennium Simulation
 - $N=656855$ Galaxies $\rightarrow N_{VL}=171,390$ Volume Limited
- A Randomly Distributed Uniform Sample is also generated with roughly the same number of points and a similar volume. $N=144,700$

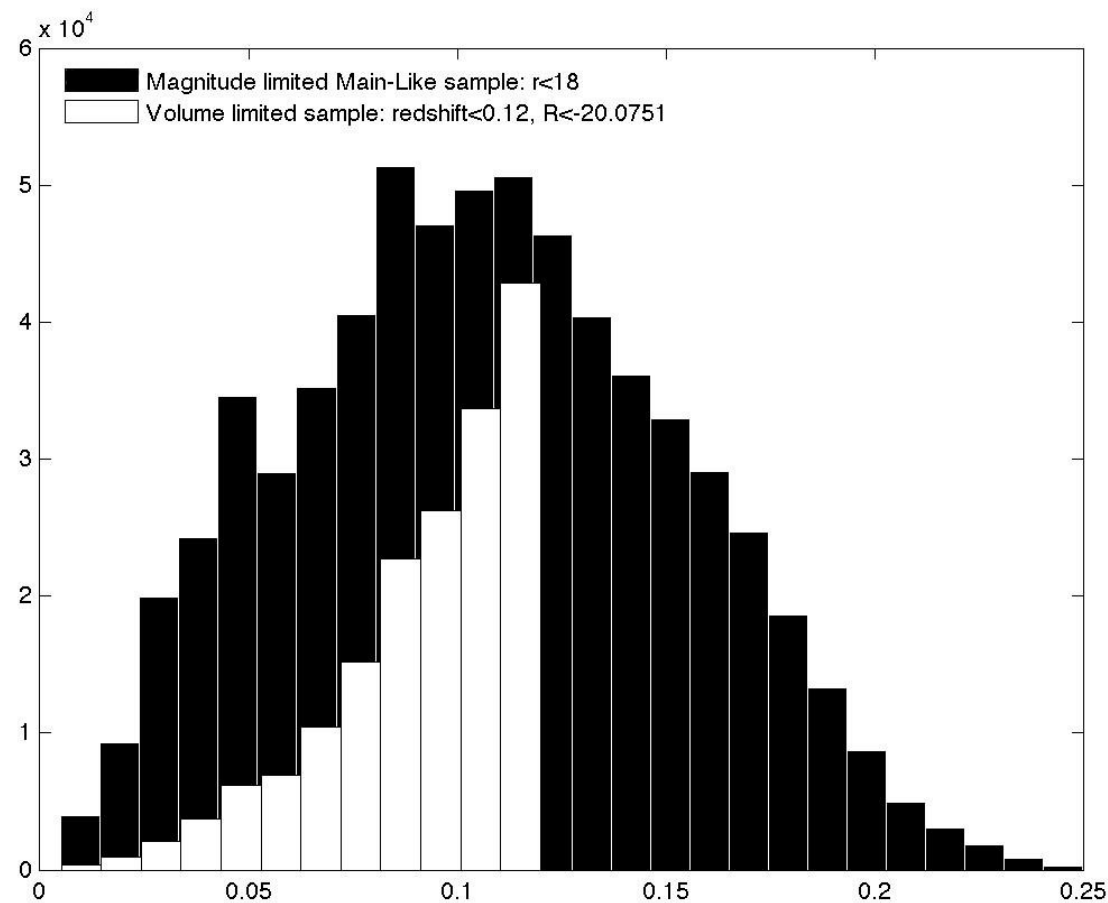
Millennium Sim Vol. Limited

Picking the volume limited sample



Millennium Sim Vol. Limited

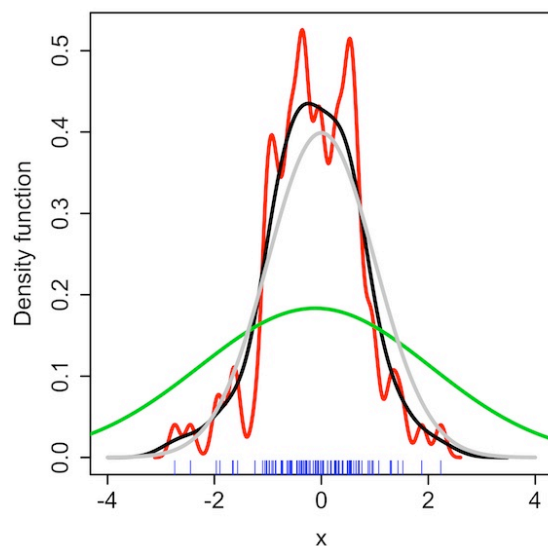
Picking the volume limited sample



Method 1: AKM

Method 1: The Adaptive Kernel Density Estimation

- 1-D example: let x_1, x_2, \dots, x_n be an independent and identically distributed random sample drawn from some unknown density f .
- We want to know the shape of this function f
- An estimate of its shape can come from the kernel density estimator. K =kernel (Gaussian is common), h =bandwidth (width of the kernel, which is a free parameter)



$$f_h(x) = \frac{1}{nh} \sum K\left(\frac{x - x_i}{h}\right) \quad K = e^{-\frac{x^2}{2h^2}}$$

Uppsala Oct 2010



Method 1: AKM

Kernel density estimate (KDE) with different bandwidths of a random sample of 100 points from a standard normal distribution.

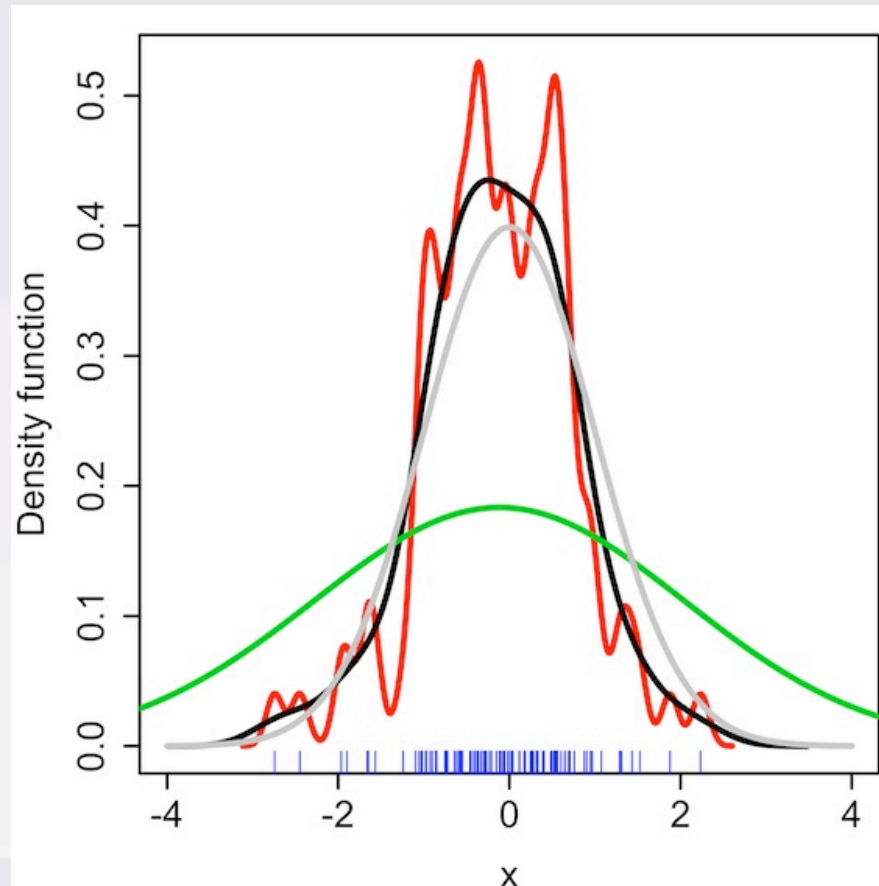
Grey: true density (standard normal)

Red: KDE with $h=0.05$.

Green: KDE with $h=2$.

Black: KDE with $h=0.337$.

$$K = e^{-\frac{x^2}{2h^2}}$$





Method 1: AKM

KDE 1-D normal: $f_h(x) = \frac{1}{nh} \sum K\left(\frac{x - x_i}{h}\right) \quad K = e^{-\frac{x^2}{2h^2}}$

Adaptive Kernel Map: $f_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h_i}\right)$

Where local bandwidth is
proportional to the sqrt
of the underlying density
fcn at the sample points

$$h_i(x_i) = \left(\frac{G}{f(x_i)}\right)^{1/2}$$

G=mean over all i of the pilot density estimate with bandwidth=h

A horizontal banner at the top of the slide featuring a collage of space-related images: a blue and white Earth horizon on the left, a grey moon in the upper center, a reddish planet in the center, a blue comet streak on the right, and a spiral galaxy in the far right.

Methods 2 & 3

Methods 2 & 3 start with Voronoi Tessellation
so lets begin there...

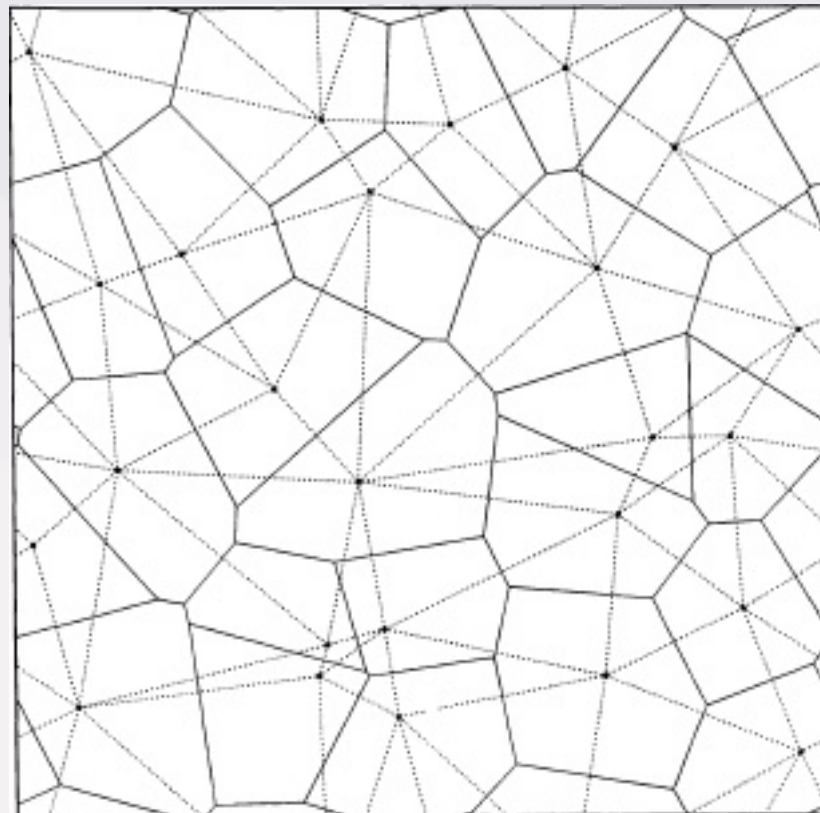


Voronoi Tessellation

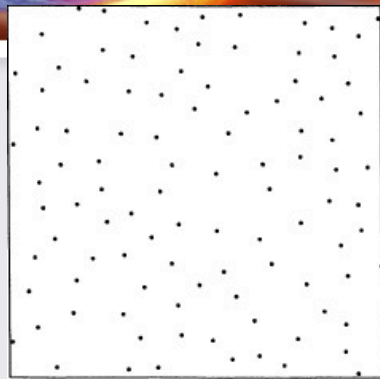
- 1) N data points generate N cells
- 2) Cells and data points are in a one-to-one correspondence
- 3) Union of all N cells is the entire data space
- 4) Intersection of any pair of cells is empty (no overlap)
- 5) Cell boundaries are flat 2-D polygons
- 6) Tessellation yields a data structure containing
 - a) Estimate of the local point density: $1/V$, V =cell volume
 - b) 3-D vector from cell centroid to data point estimates local density gradient in both magnitude and direction
 - c) Nearest-neighbor information is encoded in vertices of bounding polygons: Two cells can be adjacent in 3 ways: Do they share at least one vertex, edge or face? (Each is included in the next)

Voronoi Tessellation

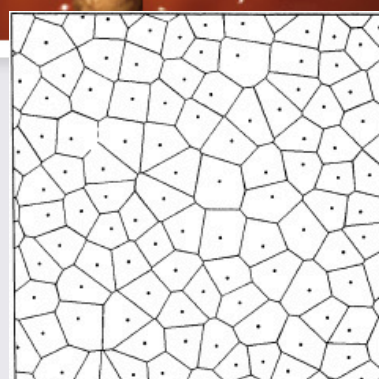
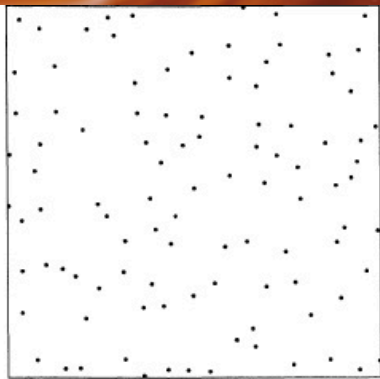
Here we have a 2-D Voronoi Tessellation (thick lines) and its corresponding Delaunay triangulation (thin lines).
from Icke and van de Weygaert 1987 (Figure 1)



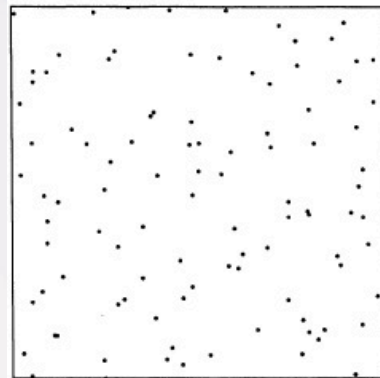
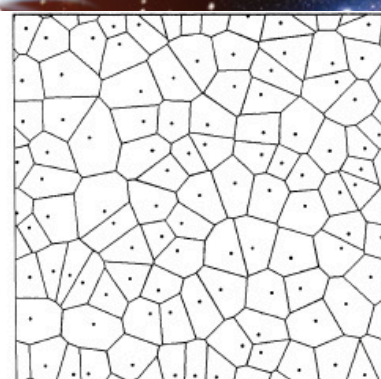
Voronoi Tessellation



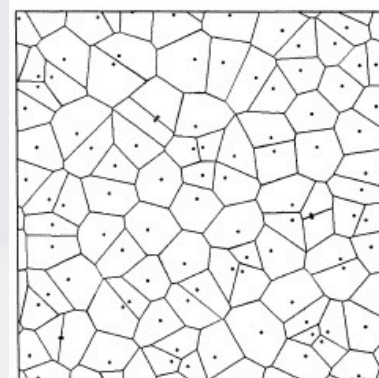
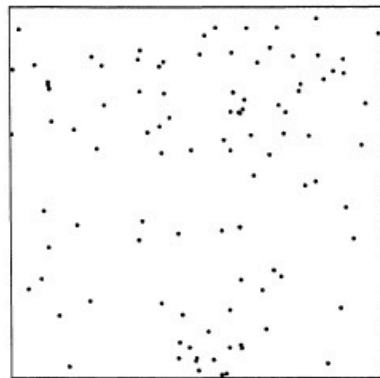
a)



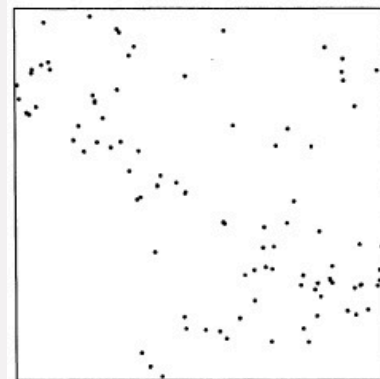
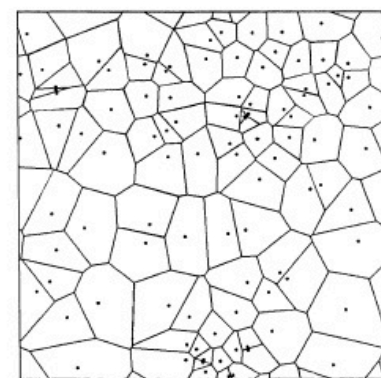
a)



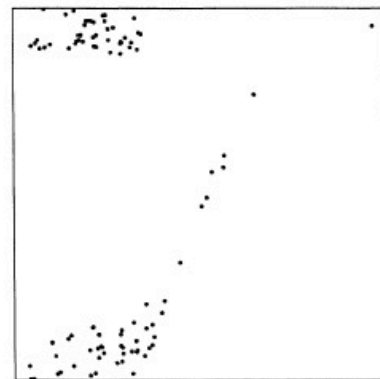
c)



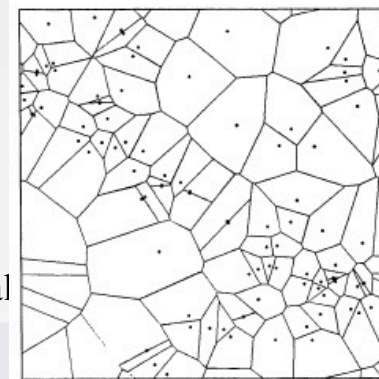
c)



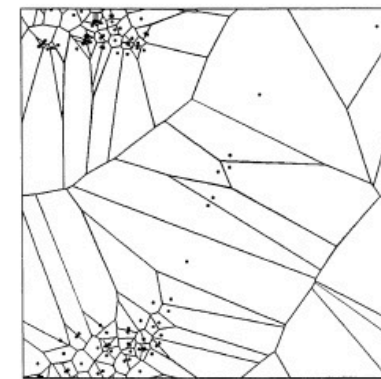
e)



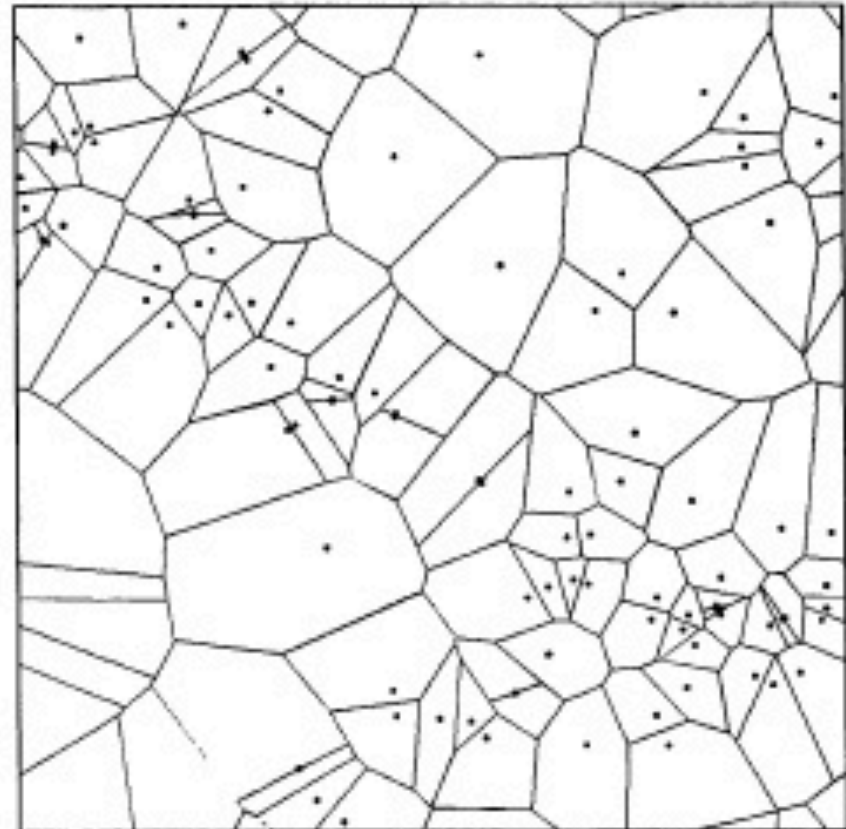
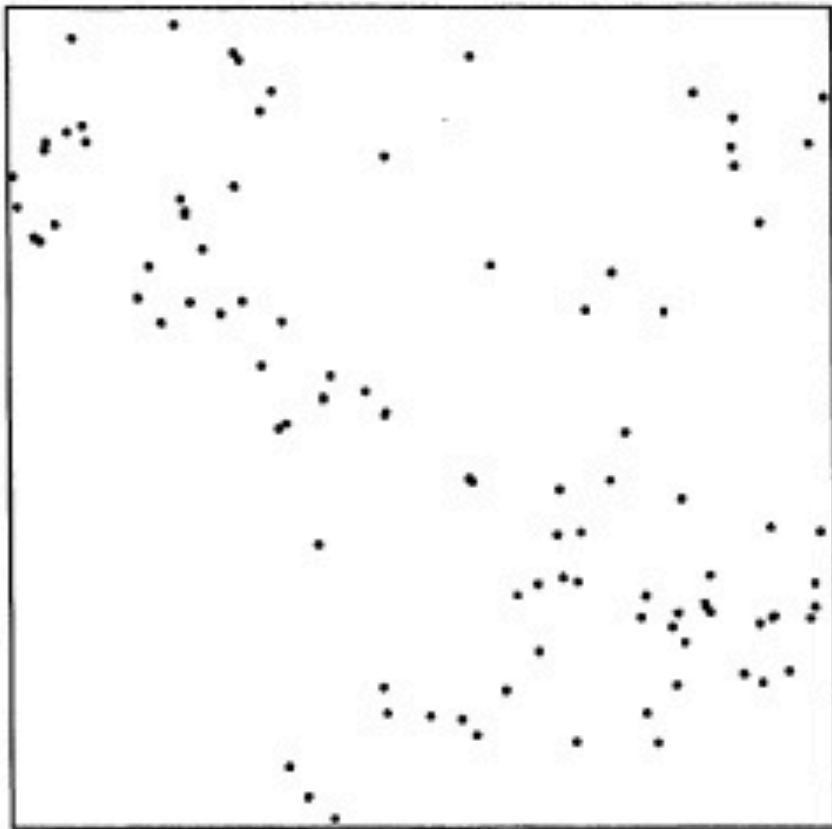
al



e)



Voronoi Tessellation



Uppsala Oct 2010





Methods 2 & 3

Now that we have our Voronoi Tessellation lets look at the methods we use to find structures.

1.) Bayesian Blocks

2.) Self-Organizing Maps



Bayesian Blocks

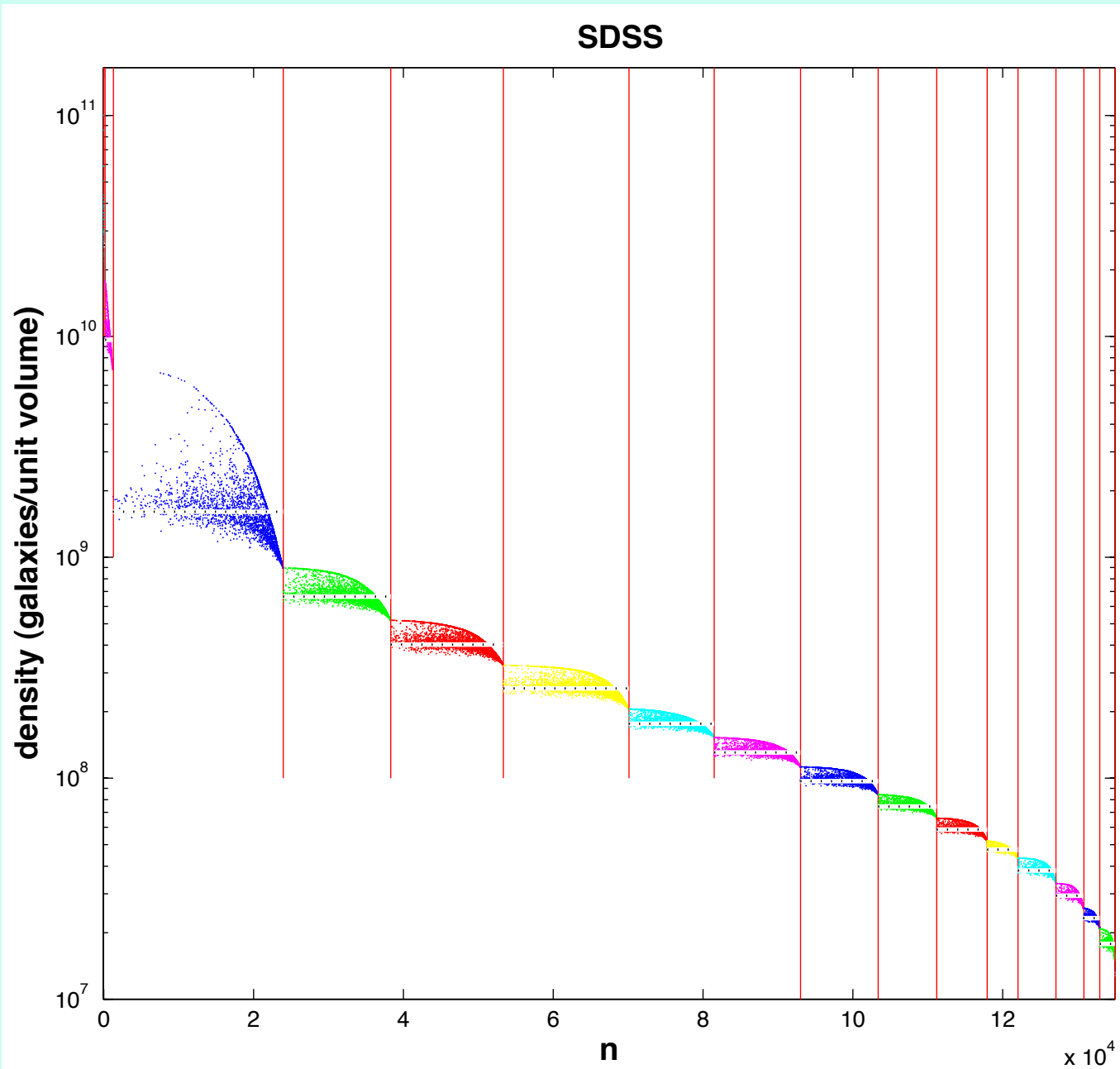
- 1) Partition data space with a set of surfaces enclosing 3-D solids
- 2) Assign a constant density to each solid = \#galaxies/volume

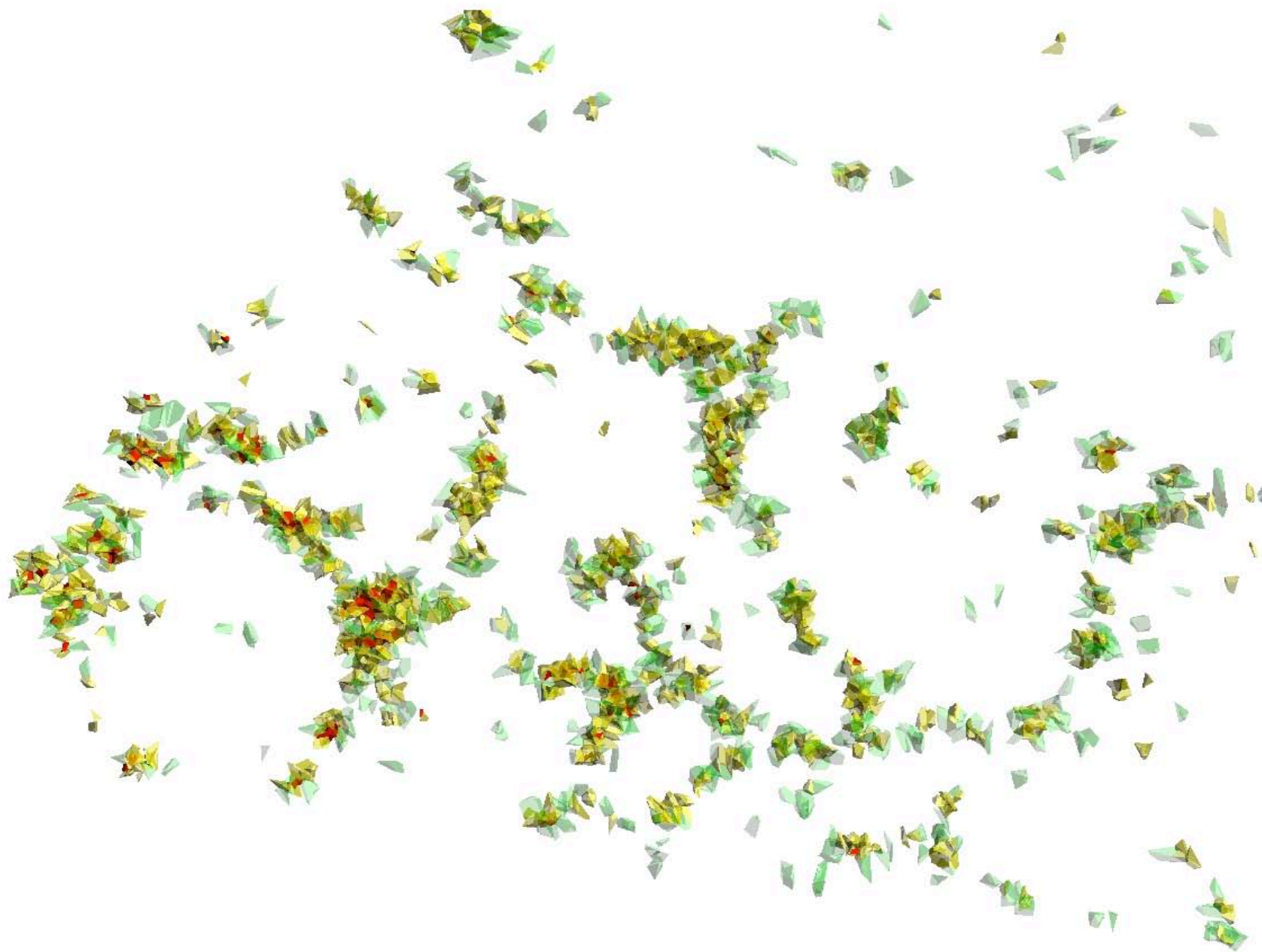
Done via an optimization procedure designed to:

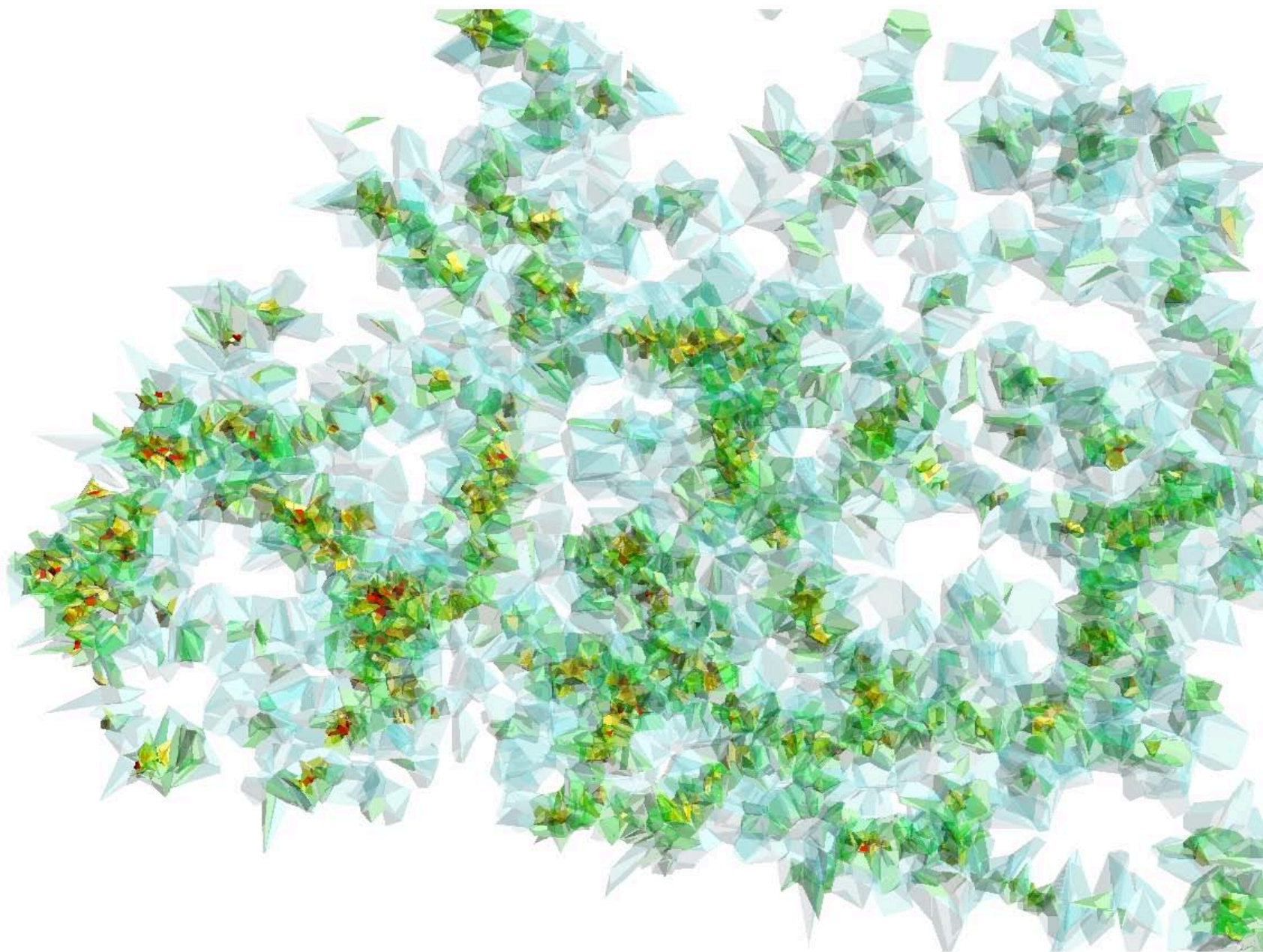
1. express spatial density variations that are real (true signal)
2. suppress statistical fluctuations that are not real (noise)

[See Scargle 1998 and Jackson et al. 2005 for the 1-D version]

Bayesian Blocks







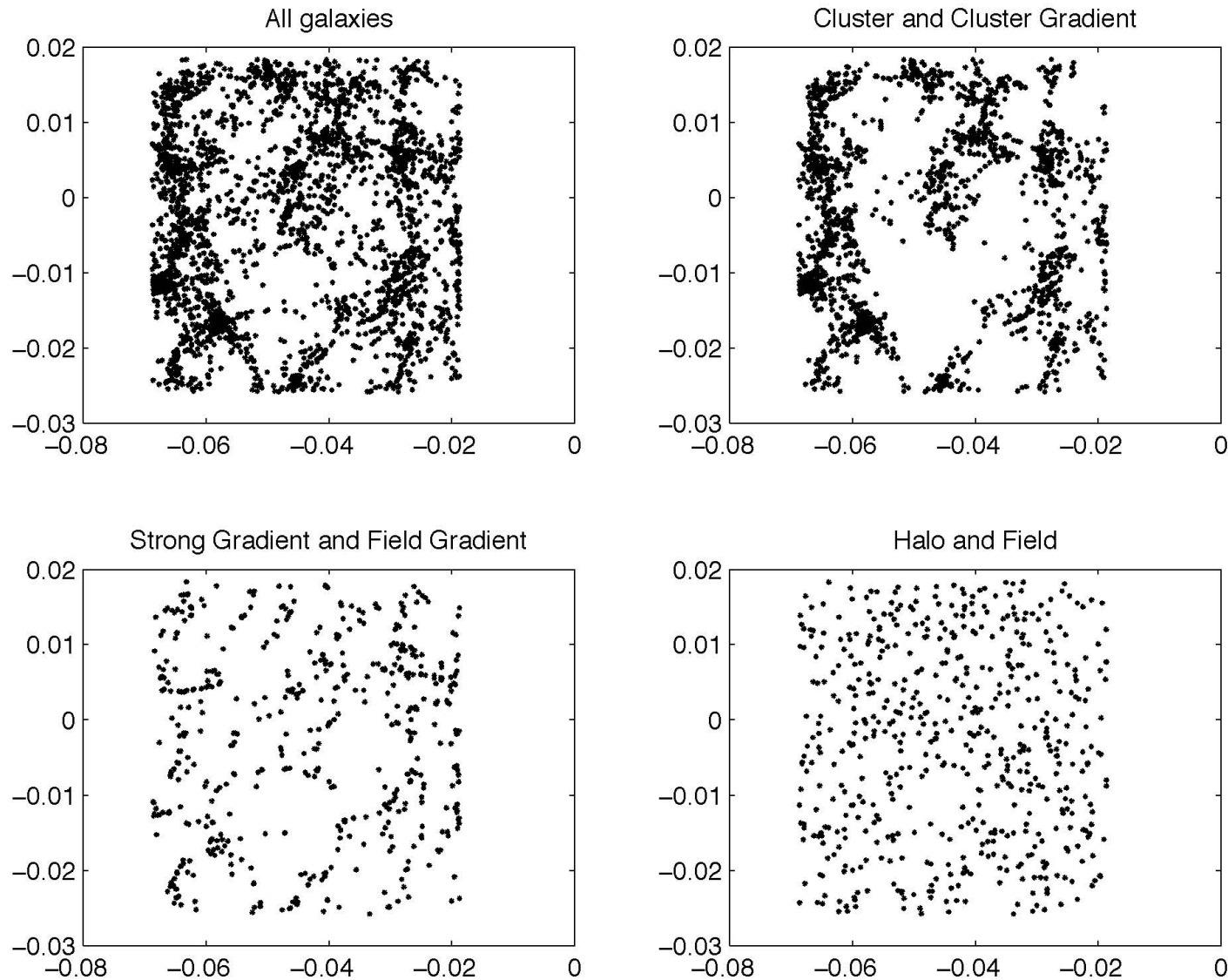


Self-Organizing Maps

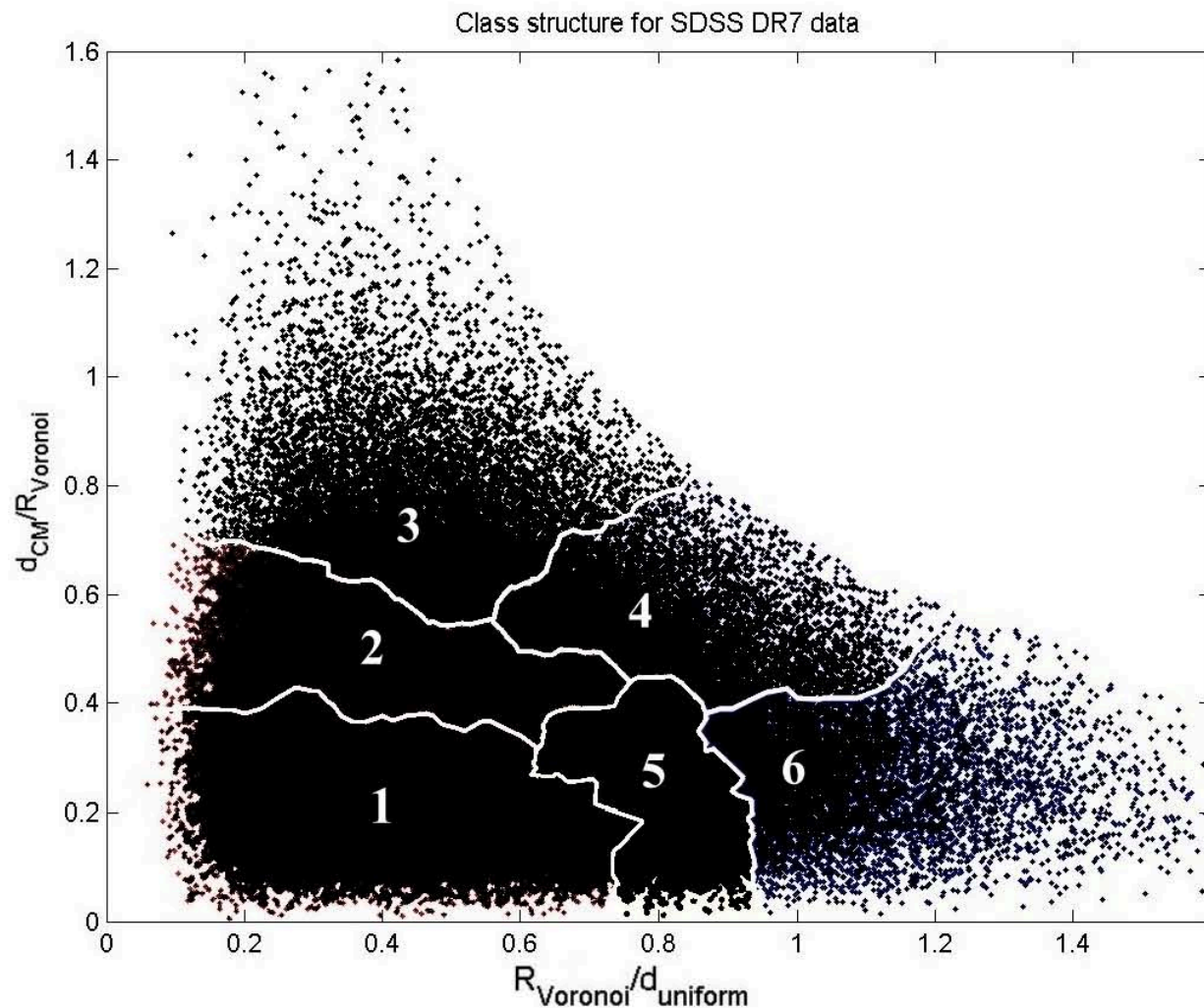
- 1) Map points from a N-Dim data space into an array of cells of principle elements (PE) in a classification space of reduced dimensionality (2-D here)
- 2) Designed (as much as possible) to reproduce the topological structure of the input distribution

Attempts to map adjacent clusters in the input space into adjacent adjacent blocks of contiguous PEs in the output space

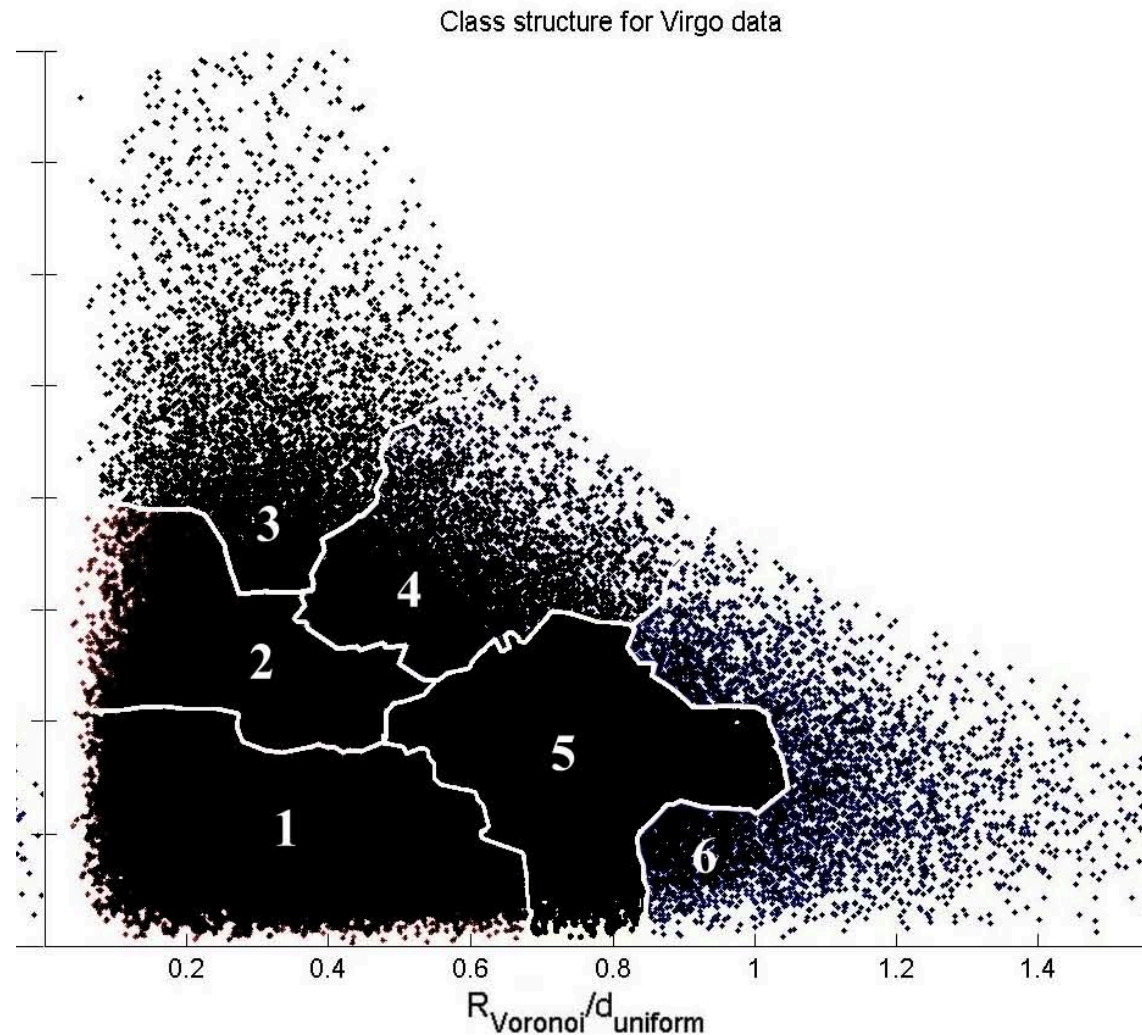
Locations in SOM phase space of types of galaxies identified by the SOM



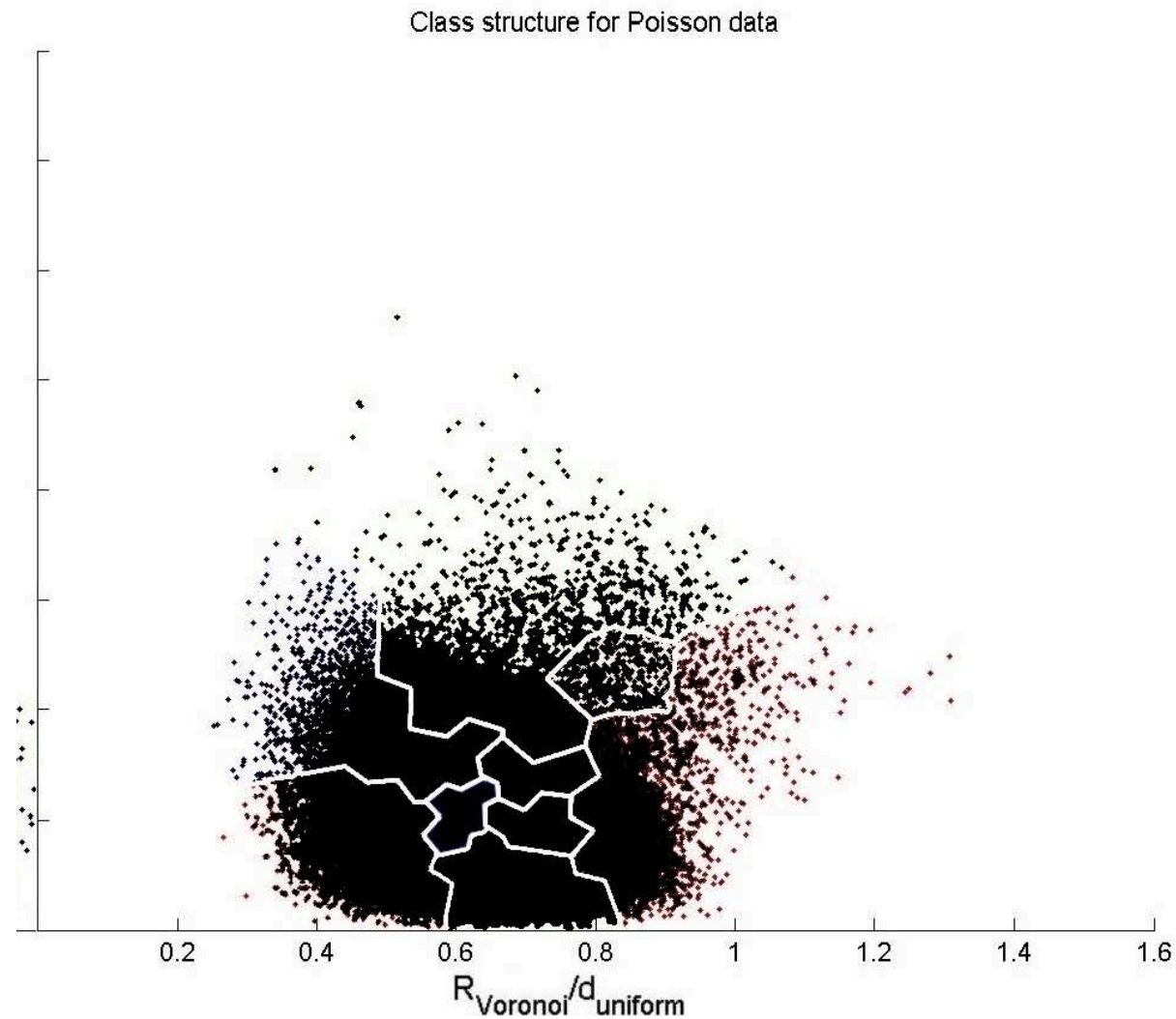
Locations in neighbor-distance/cell-vol space of galaxies assigned to various SOM classes



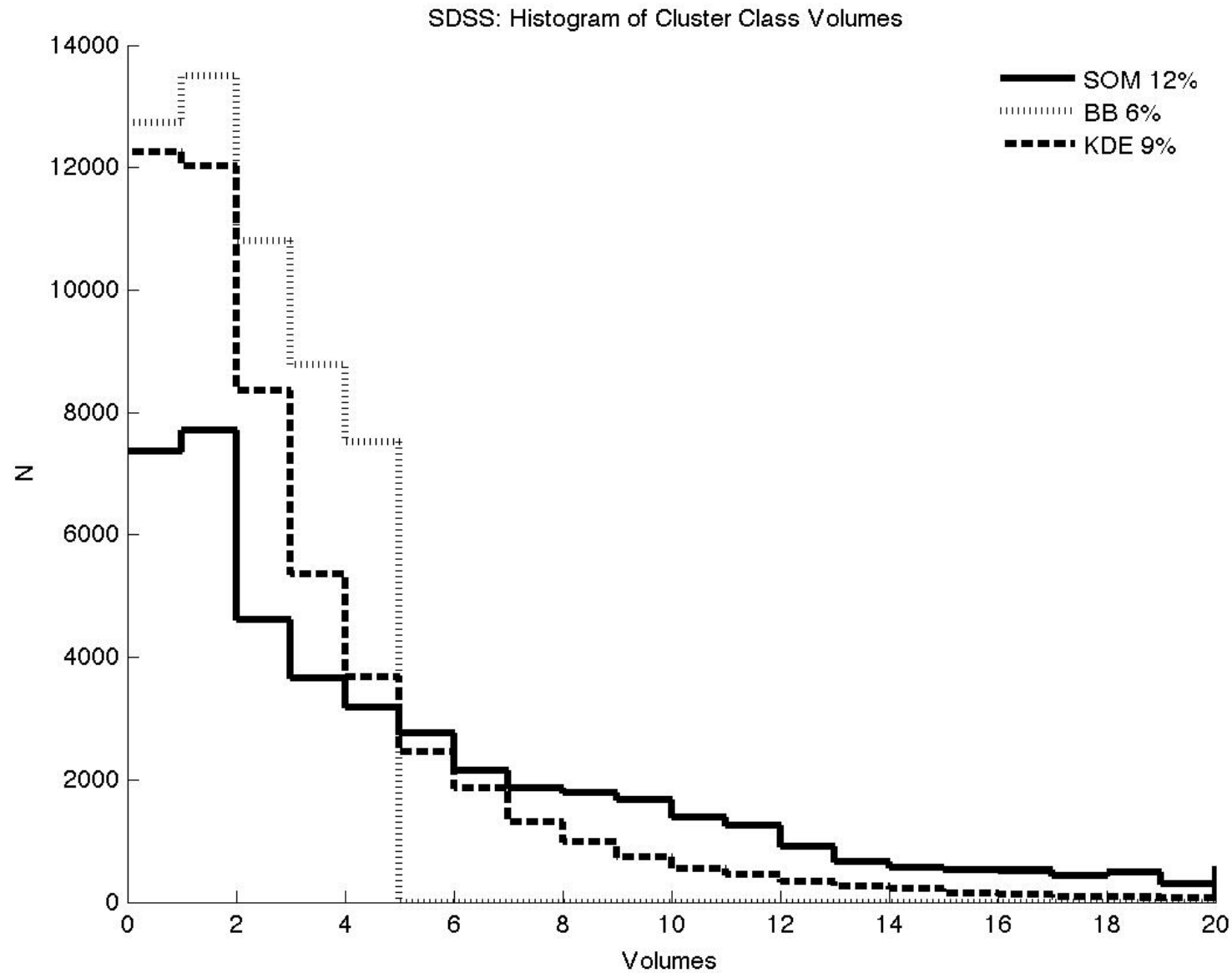
Locations in neighbor-distance/cell-vol space of galaxies assigned to various SOM classes



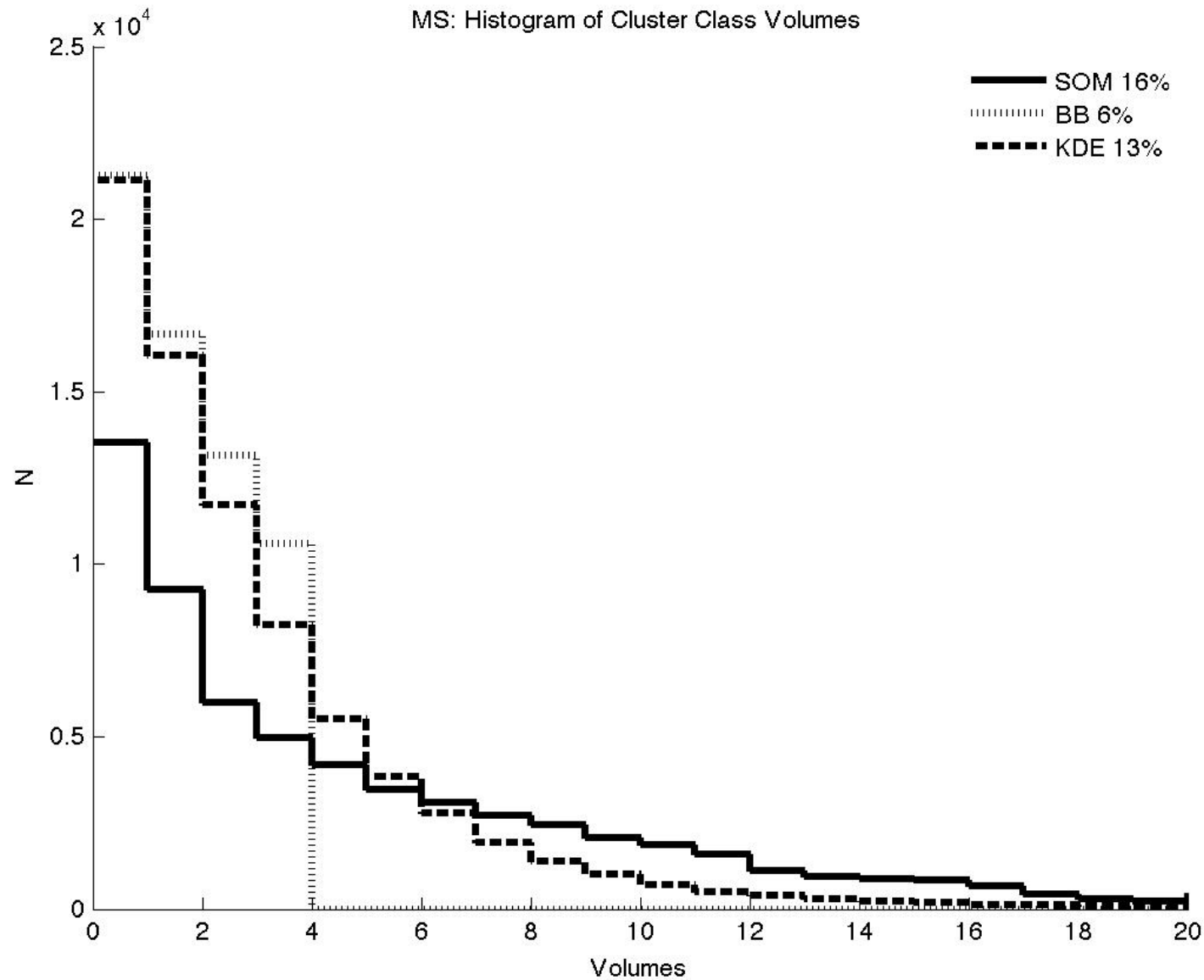
Locations in neighbor-distance/cell-vol space of galaxies assigned to various SOM classes



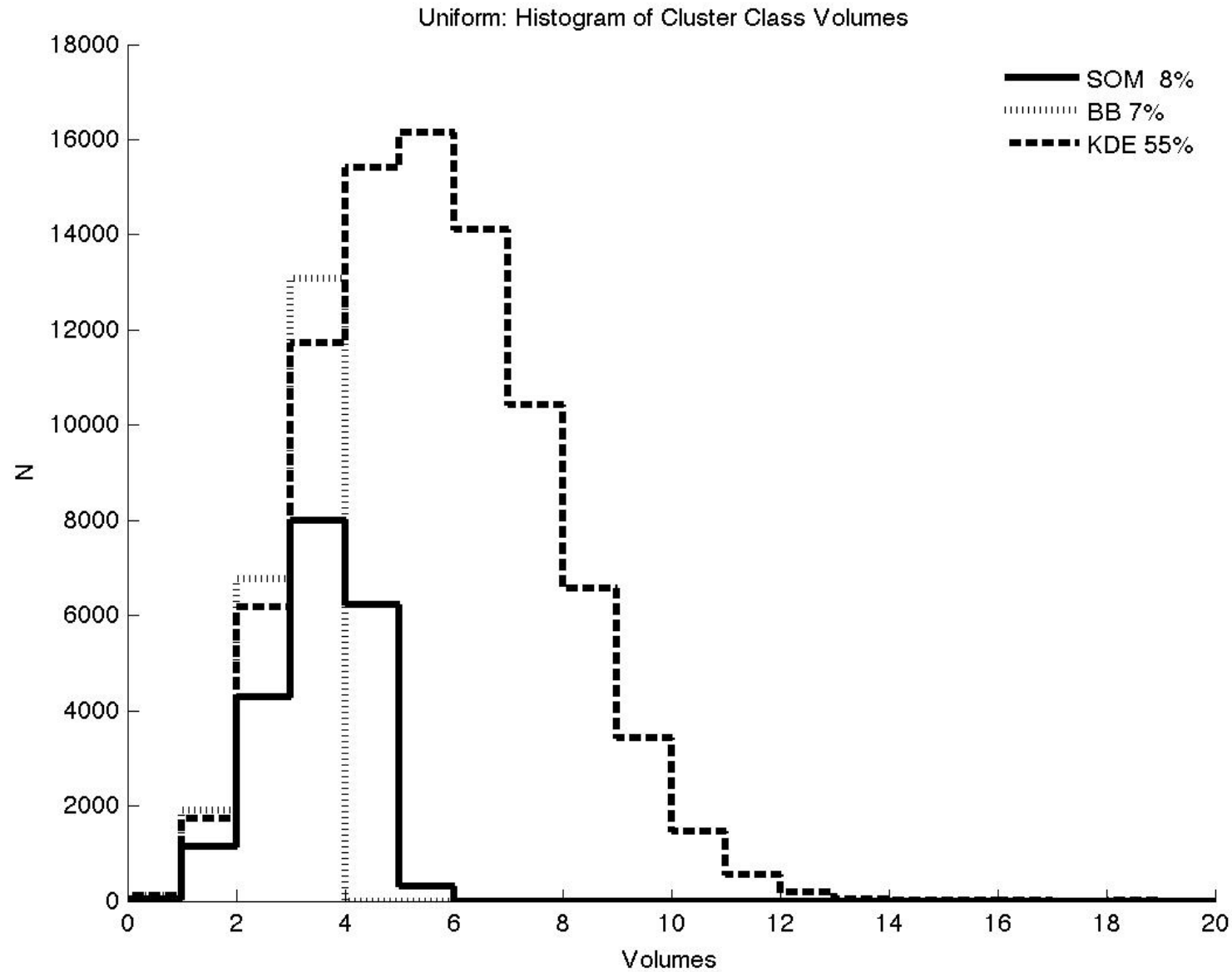
Volume Histogram for 3 methods on SDSS



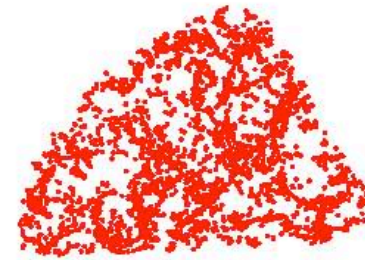
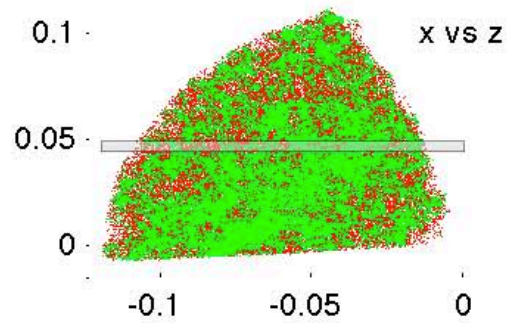
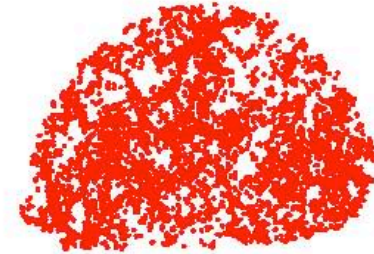
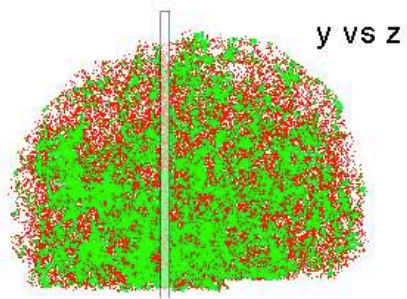
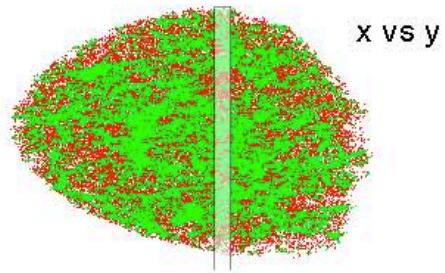
Volume Histogram for 3 methods on SDSS



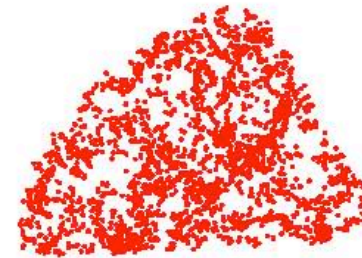
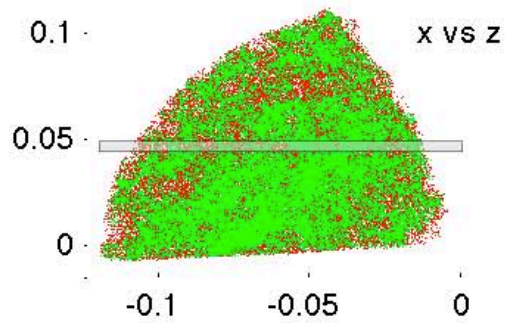
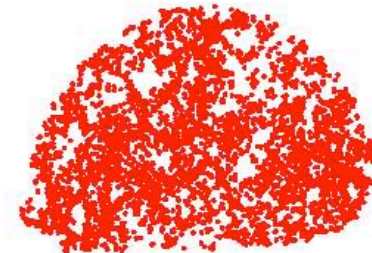
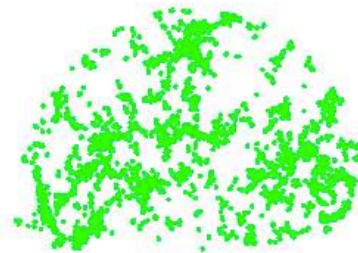
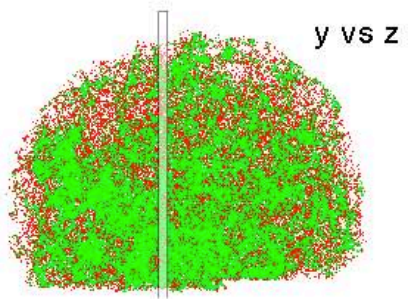
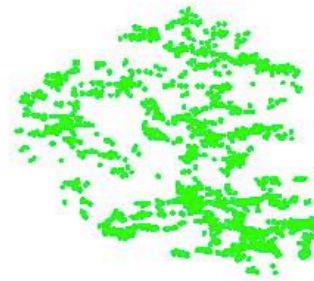
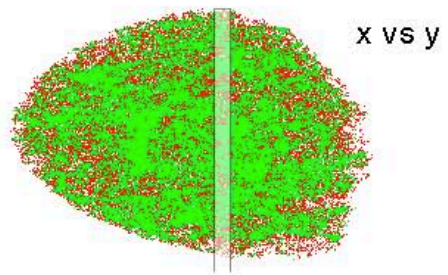
Volume Histogram for 3 methods on SDSS



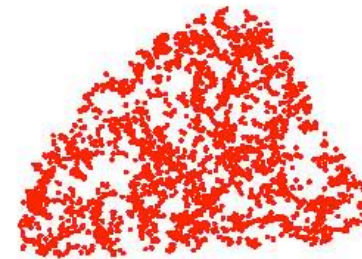
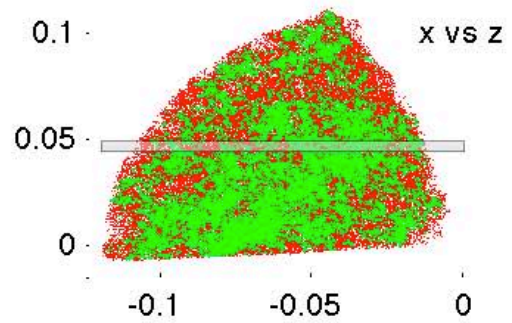
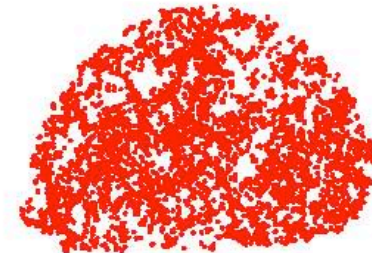
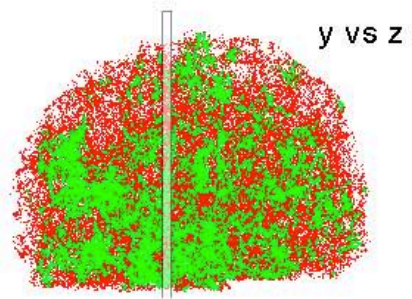
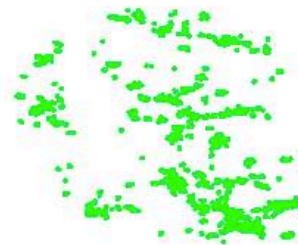
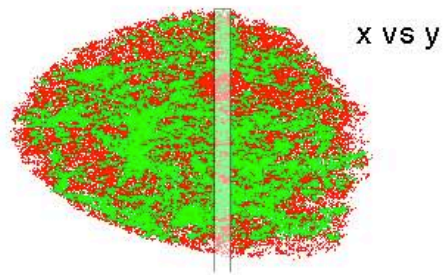
SDSS: BB Cluster Classes



SDSS: SOM Cluster Classes



SDSS: KDE Cluster Classes



SDSS: SOM Cluster Class



SDSS: BB Cluster Classes



SDSS: KDE Cluster Classes



MS: SOM Cluster Class



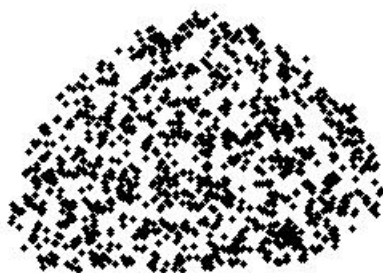
MS: BB Cluster Classes



MS: KDE Cluster Classes



Uniform: SOM Cluster Class



Uniform: BB Cluster Classes



Uniform: KDE Cluster Classes





Conclusions?

- Multi-scale structure in SDSS and Millennium Simulation have same character
- SDSS and MS are qualitatively and quantitatively different from Poisson
- BB and SOM provide similar representations of high-density structures in the SDSS and MS data
- KDE is similar to BB and SOM, but is not consistent in identifying the same high-density structures
- Poisson distribution proved a challenge for all three methods – as it should since there is no structure.



Future?

- Catalog of multi-scale structures in SDSS & MS:
 - Clusters of galaxies, Filaments, Voids
- Comparisons with other cluster and void finders
 - Dynamical Quantum Clustering
 - Watershed Void Finder, BCG, C4, etc...
- Environmental correlations with type and color
- Paper II: Catalog which anyone can use for any algorithm – easier to make comparisons between methods!!